Covid-19 South India
application of statistical methods for a predictive model

Azza M. N. Abdalghani
Milton Nicolás Plasencia Palacios
Anna Spagnolo

Department of Mathematics and Geosciences, University of Trieste

Statistical Methods for Data Science course
Final Project ∼ July 2020

Table of Contents

1. Problem Statement
2. Data Overview
3. Models
4. Results

Goal of Our Project

- Building a model to predict the **daily** affected cases of COVID-19 on the upcoming **12 days** for each state in **South of India**.
- Foucsing on South states : Tamil Nadu, Kerala, Karantaka, Telangana, Andhra Pradesh, Puducherry and Andaman and Nicobar Islands.

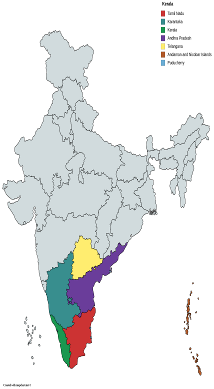


Table of Contents

1. Problem Statement
2. Data Overview
3. Models
4. Results

COVID-19 In India

- Reported on 30 January 2020, originated from China.
- Largest number of confirmed cases in Asia, and third highest number of confirmed cases in the world.
- Breaching 1,000,000 confirmed cases on 17 July 2020.
- 4 phases of lockdown.

Data Description

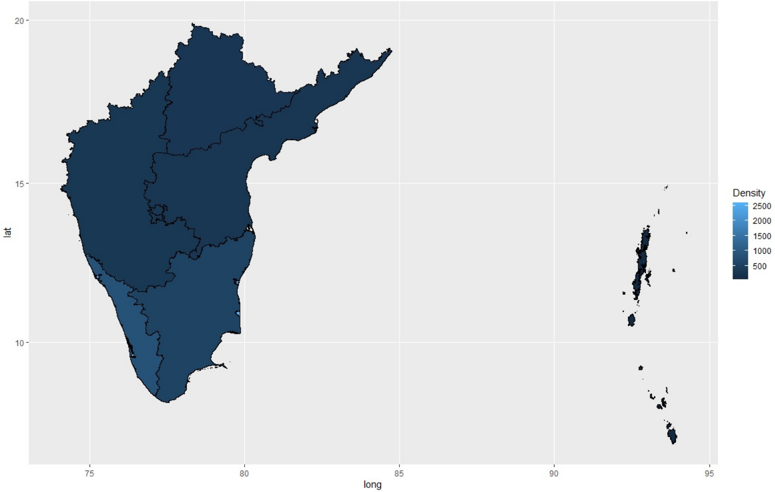
- Data Source : COVID-19 in India dataset from Kaggle.

```
COVID-19 in India
 ├── AgeGroupDetails.csv
 ├── HospitalBedsIndia.csv
 ├── ICMRTTestingLabs.csv
 ├── IndividualDetails.csv
 ├── StatewiseTestingDetails.csv
 ├── covid_19_india.csv
 └── population_india_census2011.csv
```

- Contain statewise information, personal data for the affected people and medical facilities at each state.

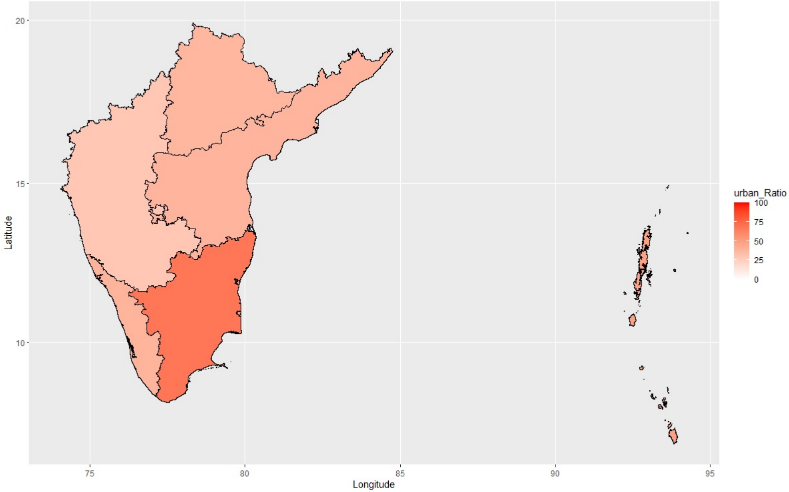
Data Visualization

* Population density per each state in South India.



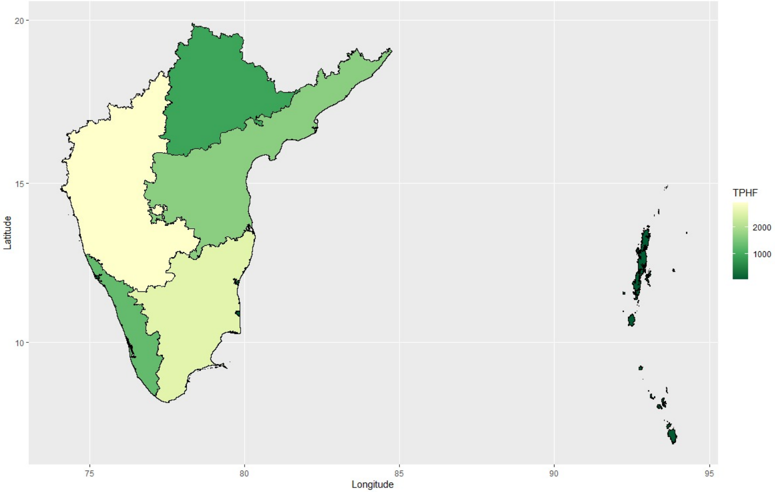
Data Visualization

* Urban Ratio per each state in South India.



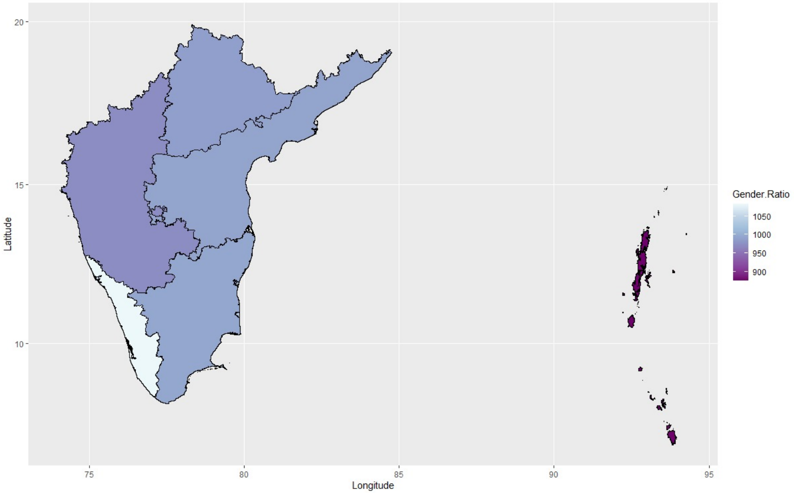
Data Visualization

* Total Public Health Facilities per each state in South India.



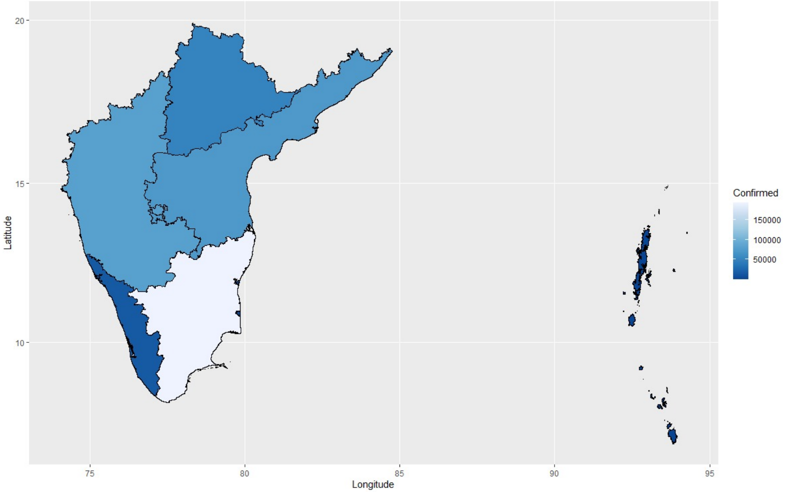
Data Visualization

* Gender Ratio per each state in South India.



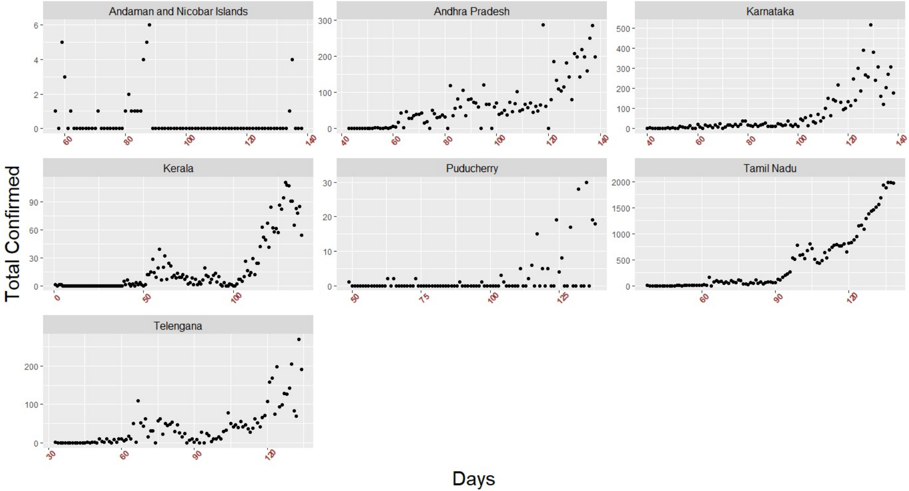
Data Visualization

* Total Confirmed Cases per each state in South India until 24/7/2020.



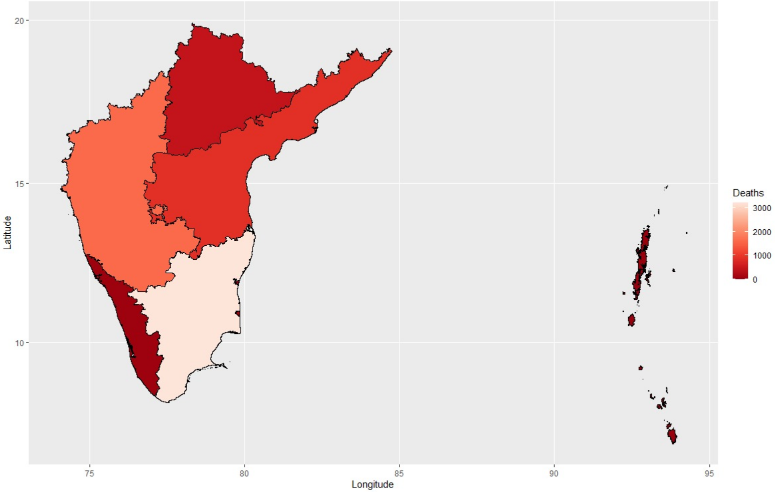
Data Visualization

* Daily Confirmed Cases per each state in South India until 24/7/2020.



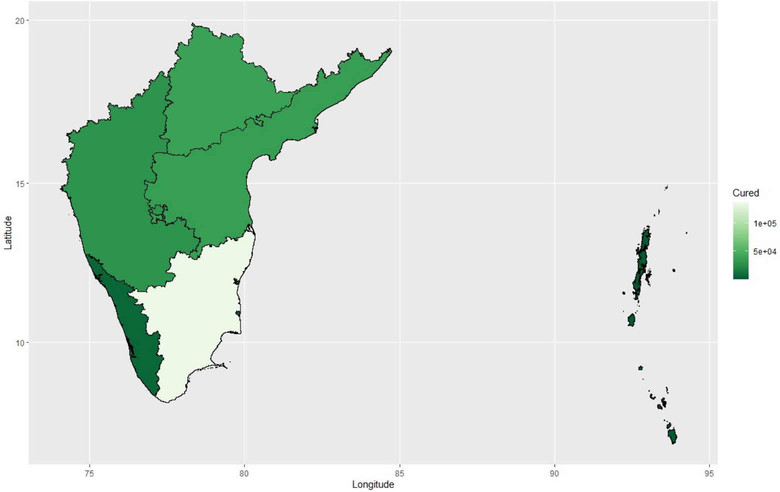
Data Visualization

* Total Death Cases per each state in South India until 24/7/2020.



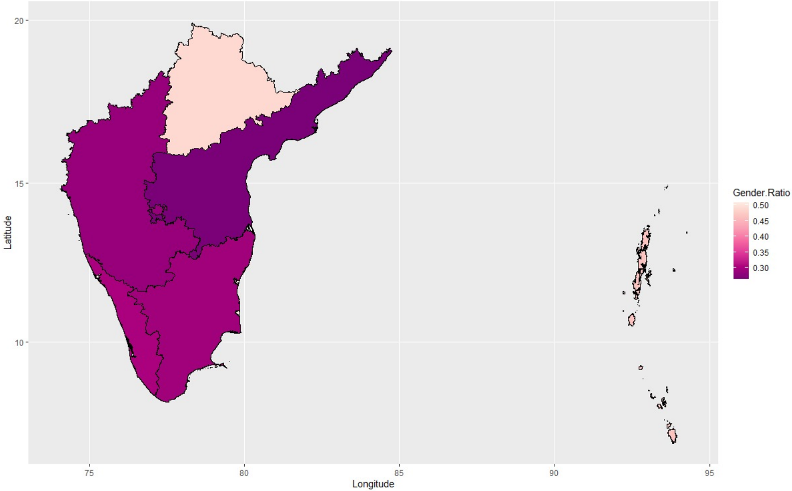
Data Visualization

* Total Cured Cases per each state in South India until 24/7/2020.



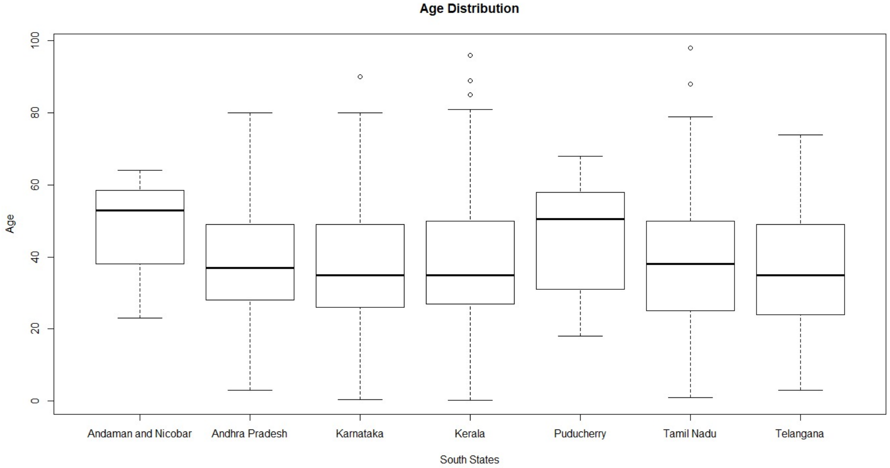
Data Visualization

* Female ratio in Confirmed cases.



Data Visualization

* Age distribution among Confirmed cases for each state.



Data Preprocessing

* Dealing with **NA** values at **(sex, age, daily_swabs)** columns :

- **sex** : **57%** of values are NA, instead we used the given ratio of infected females to infected males.
- **age** : around **70%** of values are NA, instead we used the **median** age for each state of south India.
- **daily_swabs** : **drop** NA values.

* Merge columns of interest [times, Date, daily_confirmed, age, sex, rural, lock_down, daily_swabs] to new dataset.

* Splitting dataset into **80%train** set and **20%test** datasets.

Table of Contents

1. Problem Statement
2. Data Overview
3. Models
4. Results

Methodology

- We decided to use 3 different types of model (GLM, RF and GAM) and we tried not to overfit the models (only meaningful covariates)
- Bottom up strategy
- We used a train/test split: we divided the dataset in a training set to build the model on, and a test set to be used for predictions
- For what concerns the errors, we used both the MAE and the RMSE

Generalized Linear Models (GLM)

- Extension of linear models
- Mean and linear predictor related by the link function

$$g(E(Y_i)) = g(\mu_i) = \eta_i = {x_i}^T \beta$$

- Importance of the exponential family

$$f(Y;\theta,\phi) = \exp\{\frac{Y\theta - b(\theta)}{\phi} + c(Y,\phi)\}$$

Our choices

A quasi-likelihood model is a semi-parametric model with the following released assumptions:

- $g(\mu_i) = g(E(Y_i)) = \eta_i, i = 1, \dots, n,$
- $var(Y_i) = \phi V(\mu_i), i = 1, \dots, n,$
- $cov(Y_i, Y_j) = 0$, if $i \neq j$.

Using the **quasipoisson** family, we have the same variance function of a Poisson and the canonical link of a Poisson.

We also use the **negative binomial**, which is a more flexible alternative to Poisson model. It is not a proper GLM, since it does not belong to the exponential family.

Generalized Additive Models (GAM)

They belong to a class of nonlinear models: semi-parametric regression models

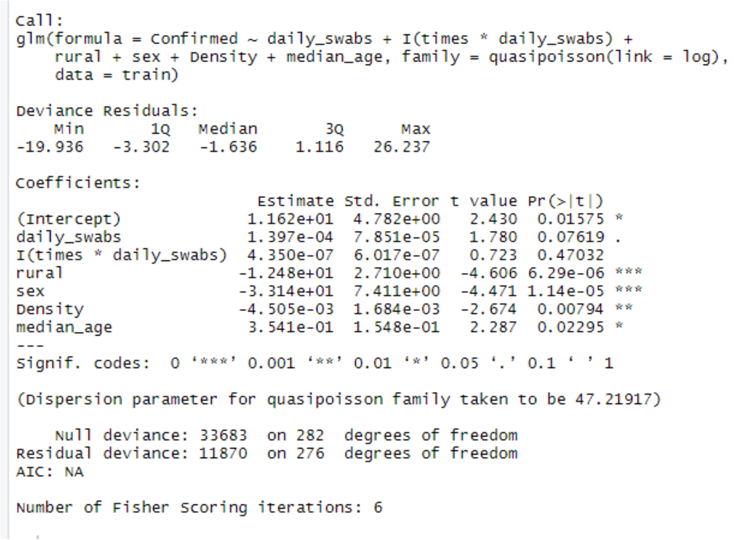
$$y_i = \beta_0 + \sum_{k=1}^{K} b_k B_k(z_i) + \text{other variables} + \varepsilon_i \quad .$$

- They can have more than one nonlinear term, which enter the specification in an **additive** way
- The response is generalized (binary or count responses are handled by a link function)

Random Forest (RF, Breiman)

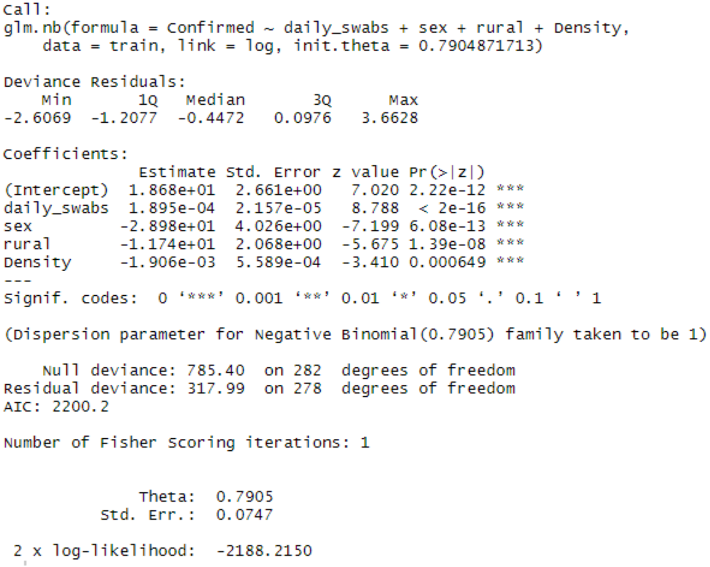
- Ensemble learning method for classification
- Used to combine a multitude of trees
- At each split only some features are considered, in order to de-correlate the trees
- Refinement of bagging: de-correlate trees and reduce variance

Quasipoisson



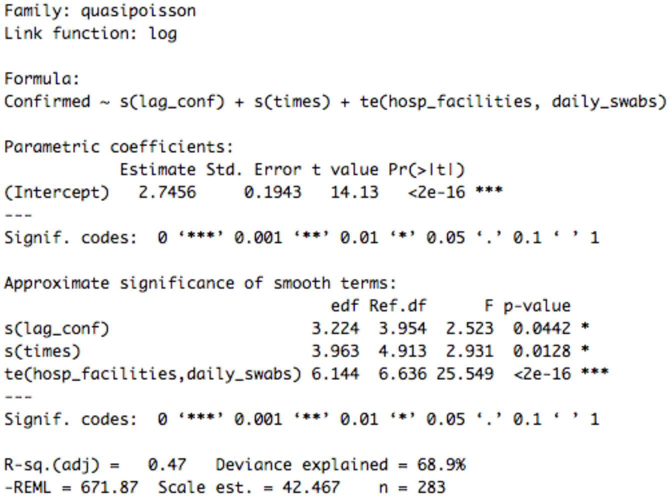
```
Call:
glm(formula = Confirmed ~ daily_swabs + I(times * daily_swabs) +
    rural + sex + Density + median_age, family = quasipoisson(link = log),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-19.936   -3.302   -1.636    1.116   26.237

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.162e+01  4.782e+00   2.430  0.01575 *
daily_swabs             1.397e-04  7.851e-05   1.780  0.07619 .
I(times * daily_swabs)  4.350e-07  6.017e-07   0.723  0.47032
rural                  -1.248e+01  2.710e+00  -4.606 6.29e-06 ***
sex                    -3.314e+01  7.411e+00  -4.471 1.14e-05 ***
Density                -4.505e-03  1.684e-03  -2.674  0.00794 **
median_age              3.541e-01  1.548e-01   2.287  0.02295 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 47.21917)

    Null deviance: 33683  on 282  degrees of freedom
Residual deviance: 11870  on 276  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Negative Binomial



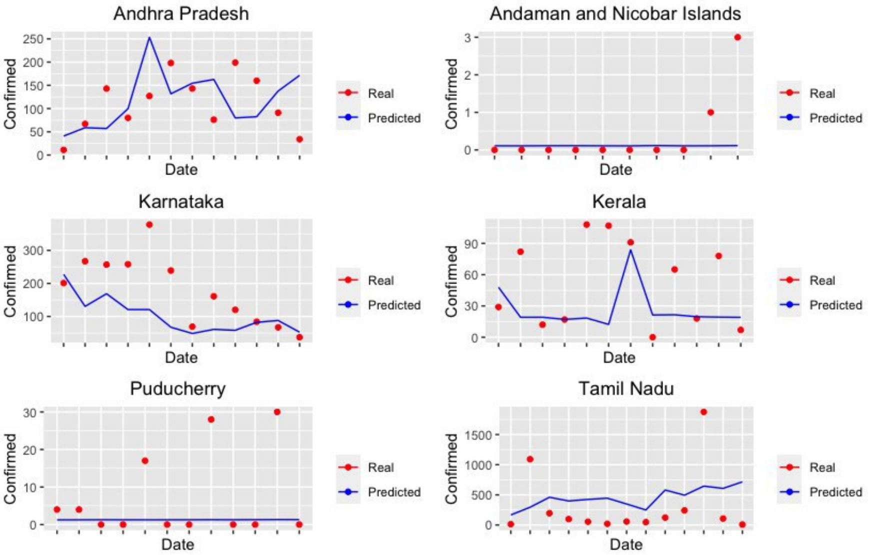
```
Call:
glm.nb(formula = Confirmed ~ daily_swabs + sex + rural + Density,
    data = train, link = log, init.theta = 0.7904871713)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6069  -1.2077  -0.4472   0.0976   3.6628

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.868e+01  2.661e+00   7.020 2.22e-12 ***
daily_swabs  1.895e-04  2.157e-05   8.788  < 2e-16 ***
sex         -2.898e+01  4.026e+00  -7.199 6.08e-13 ***
rural       -1.174e+01  2.068e+00  -5.675 1.39e-08 ***
Density     -1.906e-03  5.589e-04  -3.410 0.000649 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.7905) family taken to be 1)

    Null deviance: 785.40  on 282  degrees of freedom
Residual deviance: 317.99  on 278  degrees of freedom
AIC: 2200.2

Number of Fisher Scoring iterations: 1


              Theta:  0.7905
          Std. Err.:  0.0747

 2 x log-likelihood:  -2188.2150
```

GAM

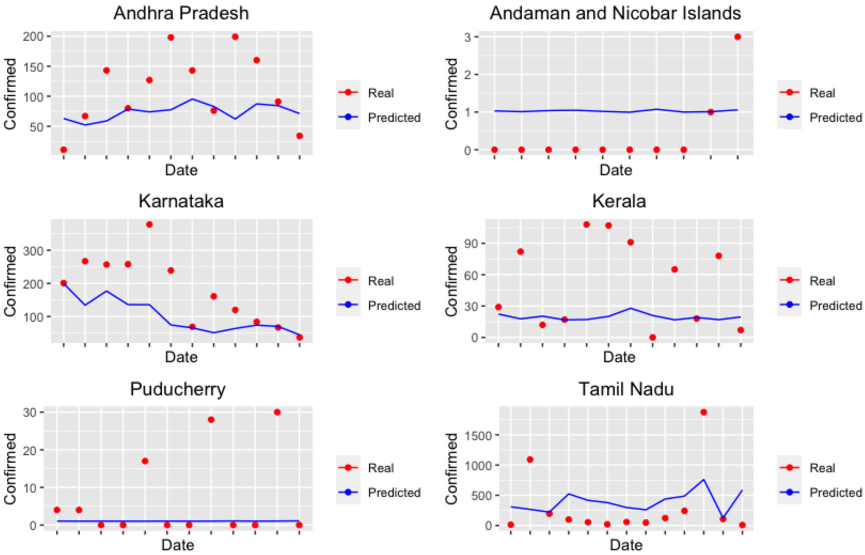


```
Family: quasipoisson
Link function: log

Formula:
Confirmed ~ s(lag_conf) + s(times) + te(hosp_facilities, daily_swabs)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7456     0.1943   14.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                edf Ref.df      F p-value
s(lag_conf)                   3.224  3.954  2.523  0.0442 *
s(times)                      3.963  4.913  2.931  0.0128 *
te(hosp_facilities,daily_swabs) 6.144  6.636 25.549  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.47   Deviance explained = 68.9%
-REML = 671.87  Scale est. = 42.467    n = 283
```

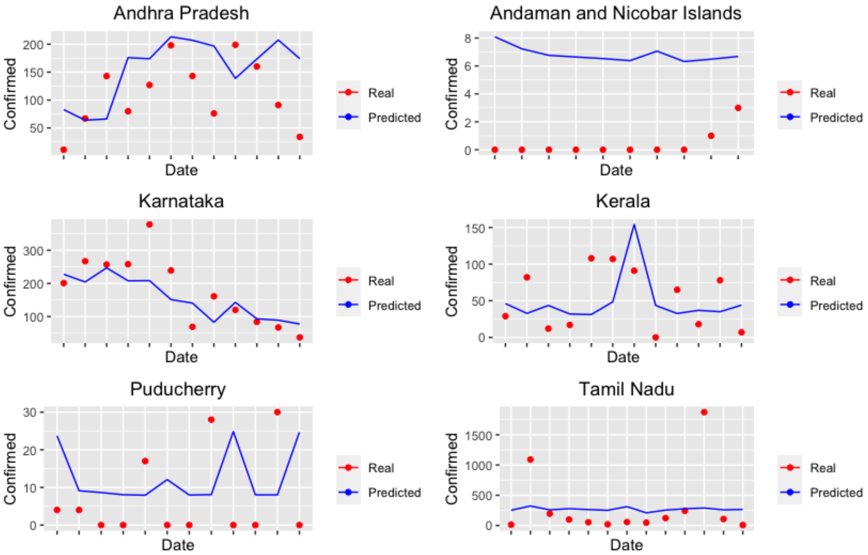
GLM predictions



GAM predictions

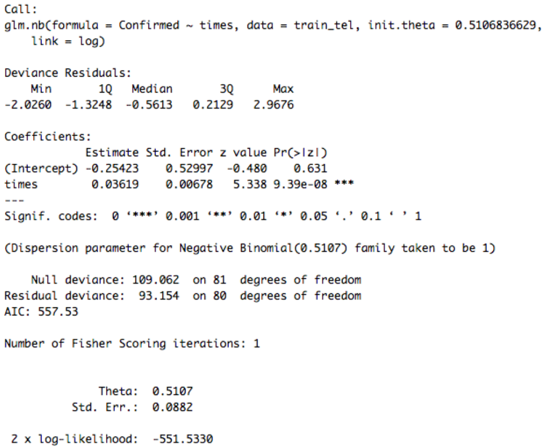


RF predictions



Telangana (Summary GLM)

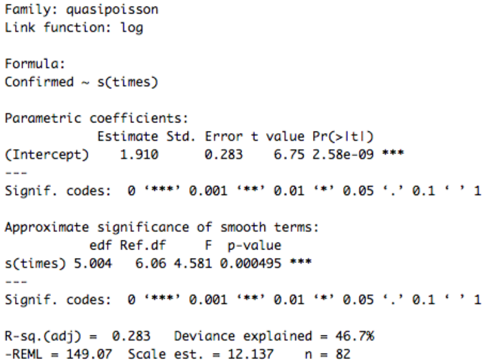
The lack of observations of the variable "Daily swabs" did not allow us to build a model that includes all the states.



```
Call:
glm.nb(formula = Confirmed ~ times, data = train_tel, init.theta = 0.5106836629,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0260  -1.3248  -0.5613   0.2129   2.9676

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.25423    0.52997  -0.480    0.631
times        0.03619    0.00678   5.338 9.39e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.5107) family taken to be 1)

    Null deviance: 109.062  on 81  degrees of freedom
Residual deviance:  93.154  on 80  degrees of freedom
AIC: 557.53

Number of Fisher Scoring iterations: 1


              Theta:  0.5107
          Std. Err.:  0.0882

 2 x log-likelihood:  -551.5330
```

Telangana (Summary GAM)

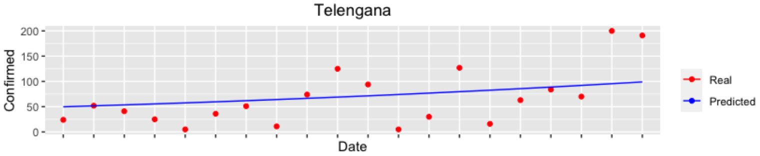
The lack of observations of the variable ”Daily swabs” did not allow us to build a model that includes all the states.

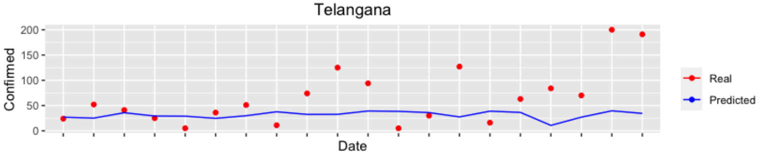


```
Family: quasipoisson
Link function: log

Formula:
Confirmed ~ s(times)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.910      0.283    6.75 2.58e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
           edf Ref.df     F  p-value
s(times) 5.004   6.06 4.581 0.000495 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.283   Deviance explained = 46.7%
-REML = 149.07  Scale est. = 12.137    n = 82
```

Telangana (Plots)

The lack of observations of the variable "Daily swabs" did not allow us to build a model that includes all the states.





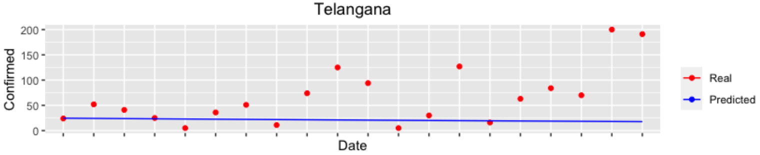


Table of Contents

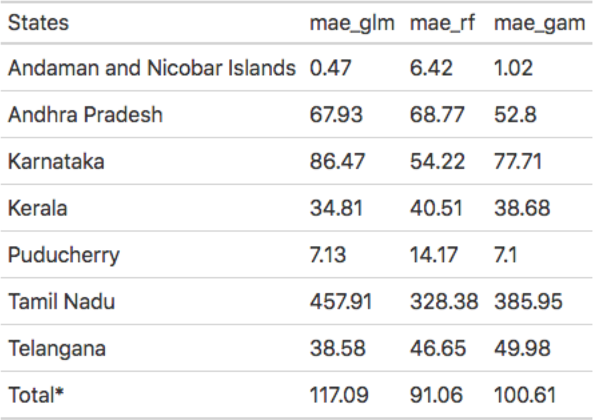
1. Problem Statement
2. Data Overview
3. Models
4. Results

Summary

Summary of the models we have used to make our predictions.

GLM	Confirmed $\sim$ daily swabs + sex + rural + density
RF	Confirmed $\sim$ times + daily swabs + sex + median age + rural + date + lockdown + hosp facilities + lag conf + density , ntree = 750, mtry = 5
GAM	Confirmed $\sim$ s(times) + te(hosp facilities, daily swabs) + s(lag conf)

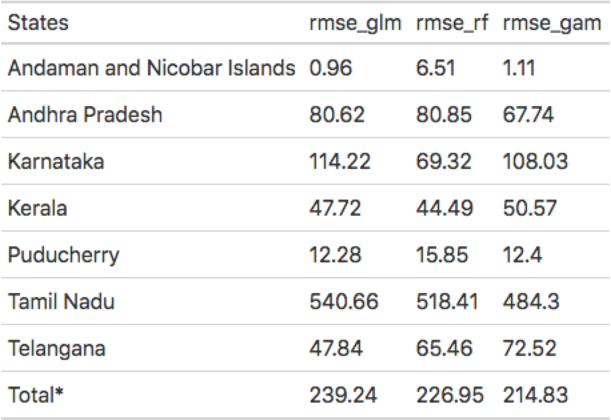
Mean Absolute Error



States	mae_glm	mae_rf	mae_gam
Andaman and Nicobar Islands	0.47	6.42	1.02
Andhra Pradesh	67.93	68.77	52.8
Karnataka	86.47	54.22	77.71
Kerala	34.81	40.51	38.68
Puducherry	7.13	14.17	7.1
Tamil Nadu	457.91	328.38	385.95
Telangana	38.58	46.65	49.98
Total*	117.09	91.06	100.61

*Total MAE has been calculated without Telangana

Rooted Mean Squared Error



States	rmse_glm	rmse_rf	rmse_gam
Andaman and Nicobar Islands	0.96	6.51	1.11
Andhra Pradesh	80.62	80.85	67.74
Karnataka	114.22	69.32	108.03
Kerala	47.72	44.49	50.57
Puducherry	12.28	15.85	12.4
Tamil Nadu	540.66	518.41	484.3
Telangana	47.84	65.46	72.52
Total*	239.24	226.95	214.83

*Total RMSE has been calculated without Telangana

Comparison & Improvements

- There's not a huge difference if we look at the error metrics.
- Random Forest has the best MAE but it is an unstable method and lack in interpretability.
- By using semiparametric regression we get the best RMSE thank to the addition of nonlinear terms.

Improvements: mainly related data collection.

Greetings

Thank you for your attention ☺